IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 14 August 2024, accepted 21 September 2024, date of publication 24 September 2024, date of current version 4 October 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3467063

RESEARCH ARTICLE

Factuality Guided Diffusion-Based Abstractive Summarization

JEONGWAN SHIN[1], HYEYOUNG PARK[1], (Member, IEEE), AND HYUN-JE SONG[2], (Member, IEEE)

[1]School of Computer Science and Engineering, Kyungpook National University, Daegu 41566, South Korea
[2]Department of Computer Science and Artificial Intelligence, Jeonbuk National University, Jeonju, Jeollabuk 54896, South Korea

Corresponding author: Hyun-Je Song (hyunje.song@jbnu.ac.kr)

This work was supported in part by the National Research Foundation of Korea (NRF) funded by the Korea Government (MSIT) under Grant 2021R1F1A1048181, and in part by the Institute of Information and Communications Technology Planning and Evaluation (IITP) funded by the Korean Government (MSIT) through Artificial Intelligence Innovation Hub under Grant 2021-0-02068.

ABSTRACT Abstractive summarization models are required to generate summaries that maintain factual consistency with the source text and exhibit high diversity to be applicable in practical applications. Existing models, which are based on pre-trained sequence-to-sequence or text diffusion approaches, generally struggle to balance these aspects, as emphasizing one typically compromises the other. To achieve both factual consistency and high diversity in summarization, this paper proposes a factuality-guided diffusion-based abstractive summarization model. This model integrates a factuality-guided module into the diffusion-based model. As the diffusion-based summarization model generates a high-diversity summary by denoising from random noise, the module guides the noise toward factual consistency with the source text. The proposed method continually guides factuality into the intermediate noise at each denoising step, thereby generating summaries that are not only consistent with the source text but also high in diversity. To guide factuality during the denoising step, this study also introduces a method for calculating the factuality based on token-level contextual matching between the source text and the intermediate noise. The effectiveness of the proposed factuality-guided summarization model is validated on three benchmark datasets, and experimental results demonstrate that the summaries generated by the proposed model are more factually consistent and diverse than those generated by baseline models.

INDEX TERMS Diffusion-based abstractive summarization, diverse text summarization, factuality-guided summarization, factually consistency in abstractive summarization.

I. INTRODUCTION

With the advancement of language models and text generation methods, there has been a significant increase in the amount of information produced in textual format. This is further augmented by automated document creation by bots, which are used in applications such as generating news articles [1], [2], shopping reviews [3], and stories [4], leading to an accelerated rate of text generation. As a result, the demand for text summarization is increasing more than ever. Text summarization is the task of generating brief and coherent summaries from the various perspectives and information contained in the source text [5], [6]. Given the critical role of summaries in conveying key information from diverse viewpoints and supporting decision-making processes, there is an increasing need to develop summarization models that not only maintain factual consistency with the source text but also generate diverse summaries.

Several models have been proposed to achieve factual consistency with the source text [7], [8], [9], [10], [11], [12] and to generate summaries with high diversity [13], [14], [15], [16], [17]. While each model demonstrates high performance in metrics tailored to its specific purpose, a trade-off between factual consistency and diversity is often observed. Table 1 presents summaries generated by the BART summarizer [6], a widely-used model for abstractive summarization, alongside

The associate editor coordinating the review of this manuscript and approving it for publication was Mohammad J. Abdel-Rahman.

TABLE 1. Summaries generated by the BART summarizer using beam search, top-k, and top-p sampling methods, alongside the autoregressive diffusion summarizer. The underline in the generated summaries indicates factual inconsistencies with the source text.

Source text	An Asiana Airlines plane overran a runway while landing at Japan 's Hiroshima Airport on Tuesday evening , prompting the airport to temporarily close , the Japanese transportation ministry said. Twenty-three people had minor injuries after Flight 162 landed at 8:05 p.m, according to fire department and ministry sources. There were 73 passengers and eight crew members – including five cabin attendants , two pilots and a maintenance official – aboard when the flight took off from South Korea 's Incheon International Airport at 6:34 p.m. local time , Asiana said in a statement late Tuesday. Authorities are investigating initial reports that the Airbus A320 may have hit an object on the runway during landing , causing damage to the rear of its body and the cover of the engine on the left wing , the ministry said. ...
Golden summary	The plane might have hit an object on the runway, the Japanese transportation ministry says. 23 people have minor injuries, officials say.
BART using beam search	0. An Asiana Airlines plane overran a runway while landing at Japan's Hiroshima Airport. Twenty-three people had minor injuries after Flight 162 landed at 8:05 p.m. 1. An Asiana Airlines plane overran a runway while landing at Japan's Hiroshima Airport on Tuesday evening. Twenty-three people had minor injuries after Flight 162 landed at 8:05 p.m.
BART using top-k	0. An Asiana Airlines flight overran the runway while landing at Japan's Hiroshima Airport Tuesday night. It is reported that 23 people have minor injuries. 1. 23 people were hurt after the crash, and Hiroshima Airport was temporarily reopen. Authorities are investigating.
BART using top-p	0. 23 people had minor injuries after Flight 162 landed at 8:05 p.m. There were 73 passengers and 8 crew members aboard the Asiana plane. 1. 23 people were injured, mainly by damage to the plane's body. Plane is owned by Airbus, which is working to gather more information, spokeswoman says.
AR-diffusion	0. Asiana says the plane overran a runway at the hiroshima airport. Authorities are investigating initial reports may have hit an object on the runway during landing. 1. Twenty-three people were minor at japan's hiroshima airport. Airbus a320 may have hit the cover of the engine on the left wing.

those generated by AR-diffusion summarizer [17]. In this table, two summaries are generated from each summarizer. For the BART summarizer, different methods such as beam search, top-p, and top-k samplings are adopted to control diversity. As shown in Table 1, while summaries generated by BART using the beam search tend to be factually consistent with source text, they exhibit low diversity. On the other hand, summaries generated by BART using the top-p and top-k samplings are more diverse than those generated with the beam search, but sometimes compromise factual consistency. Summaries from the AR-diffusion summarizer exhibit the highest level of diversity, yet one summary is factually inconsistent with the source text. These results are because models like BART do not prioritize diversity whereas diffusion-based models like AR-diffusion, which focus mainly on diversity, do not explicitly consider factual consistency. Therefore, for abstractive summarization models to achieve both high factual consistency and diversity it is crucial to consider both aspects.

This paper proposes an abstractive summarization model that aims to generate summaries that are both diverse and factually consistent. The proposed model is built upon diffusion-based text summarization [16], [17], which progressively corrupts a summary with random noise through a forward noising process, and subsequently reconstructs the summary from random Gaussian noise using multiple denoising steps in a reverse process. During the reverse process, the denoising is performed while referring to encoded representations from the source text. At inference time, this reverse process begins with sampling noise from a Gaussian distribution and iteratively denoises it to generate a summary. While the diffusion-based text summarization is beneficial for enhancing diversity, its reverse process does not explicitly ensure the factual consistency of the summaries with the source text, often resulting in summaries that are diverse yet factually inconsistent.

To address this issue, the proposed model incorporates a factuality-guided module into the reverse process as a plug-and-play component [18]. This module computes a factuality by comparing an intermediate output of the denoising step against the source text. Then, in the reverse process, the module employs information guidance [19] to ensure that the denoising not only reconstructs the summary but also maintains the factuality. With the guided module performing at each denoising step, the summarizer generates texts that are both diverse and factually consistent. It is important to note that the source text is represented as a sequence of tokens, whereas the denoising output is a representation in a continuous vector space. Thus, directly computing the factuality between these two representations is not straightforward. To tackle this, the proposed model leverages the diffusion model's embedding to embed the source text into a continuous space. The factuality is then computed as the inner product between tokens' embeddings [20].

The effectiveness of the proposed model is evaluated using three standard benchmark datasets for abstractive summarization. Experimental results demonstrate that the proposed model not only achieves high ROUGE scores but also excels in various factual consistency metrics, outperforming existing diffusion-based summarizers [16], [17], BART summarizers [6], and LLM-based summarizers [21], [22]. Furthermore, this paper adopts a large language model to evaluate the quality of the generated summaries and demonstrates that the proposed summarizer generates high-quality summaries compared to the baselines.

The main contributions of this paper are as follows:

- This is the first attempt to enhance factual consistency in diffusion-based text summarization. It introduces a method designed to balance the trade-off between diversity and factual consistency. While existing diffusion-based summarizers primarily focus on enhancing diversity and fluency, the proposed method

aims to improve factuality by incorporating an explicit factuality guide. As a result, the proposed method generates summaries that are both diverse and factually consistent.
- This paper introduces the guidance of factuality in a plug-and-play manner. As retraining diffusion-based summarization models for factuality consistency is computationally expensive, the plug-and-play approach facilitates the generation of factually improved summaries at a reduced cost.
- This paper examines the performance of the proposed model on three standard benchmark summarization datasets. Experimental results across these datasets indicate that the proposed model not only promotes high diversity but also ensures factual consistency in the generated summaries.

The rest of this paper is organized as follows. Section II reviews studies related to enhancing factuality consistency in abstractive summarization, diverse text generation, and the text diffusion model. Section III introduces the essential preliminary to the proposed model. and Section IV presents the proposed factuality guided diffusion-based summarization. The experimental settings and results are given in Section V. Finally, Section VI draws some conclusions.

II. RELATED WORK

A. ENHANCING FACTUALITY OF ABSTRACTIVE SUMMARIZATION

Recent years have seen active research in enhancing the factuality of abstractive summarization. This has led to the proposal of various approaches, including the use of external knowledge, post-editing of summarizations, and decoding stage adjustments. The first approach involves incorporating external knowledge, derived from information extractors and parsers, into abstractive summarization models. For example, Cao et al. [9] utilized an open information extractor and a dependency parser to extract triples from the source text, which are then integrated into the summarization model to generate the final summary. Similarly, Zhu et al. [11] employed a graph extracted from the source text, utilizing graph information in the summary generation process via graph attention. Li et al. [23] introduced a Transformer-based Entity Augmented Method that includes two novel modules: a sparse entity matrix following the encoder and an additional entity cross-attention layer in the decoder, aimed at seamlessly integrating entity boundary information into the summarization model. However, a primary limitation of this approach is its model-specific nature, meaning the methods are only effective in conjunction with specific summarization models.

The second approach is the post-editing correction that takes a model-generated summary and the source text and corrects the summary to ensure factual consistency with the source text. This method is model-agnostic, because it can receive any generated summary as its input. Several studies have explored this approach, such as the rewriting method [24] and span correction [25], [26], which vary in their degrees of correction. The rewriting method regenerates a new summary through an autoregressive sequence-to-sequence model, which potentially leads to significant differences from the original summary. In contrast, the span correction method makes partial modifications to the inconsistent spans of the generated summary, resulting in less variation. However, this approach often resembles extractive summarization due to its limited scope of modification. It is noteworthy that these approaches primarily focus on improving factuality, with little emphasis on enhancing diversity. The primary objective of this paper, however, is to improve both factuality and diversity in summarization generation, rather than to correct an already generated summary.

The last approach focuses on the decoding stage to select factually consistent summaries. This approach chooses the most coherent summary during decoding, independent of model specifics. Wan et al. [8] posited that beam search could explore various candidates, among which a more faithful summary might exist, even if it is not the highest-scoring according to the model. They suggested re-ranking these candidates using a faithfulness metric. Pernes et al. [7] introduced an energy-based model for re-ranking summaries based on factuality metrics. Additionally, Dixit et al. [27] developed a method that employs a ranking technique post-summary generation for comparative summary training, enhancing both the factuality and quality of the abstract summary. While decoding methods have improved factual consistency, they often reduce the diversity among the candidate summaries.

B. DIVERSE TEXT GENERATION MODELS

The generation of highly diverse texts has received significant interest across various text generation applications, including paraphrase generation [28] and response generation [29]. Research for the diverse text generation explores a range of methods: employing different decoding strategies into the decoder, incorporating randomness on the encoder [14], and utilizing reinforcement learning for decoder training [28], [30]. Specifically, decoder-side approaches have led to the development of various decoding algorithms to enhance diversity such as diverse beam search [31], top-k sampling [32], and top-p sampling [33]. While the primary goal of these studies focuses on generating diverse texts, the aspect of factual consistency is not explicitly addressed.

C. TEXT DIFFUSION MODEL

Diffusion models [34] have shown remarkable efficacy in image generation, which has led to recent investigations into the extension of diffusion models to text domains [35]. Diffusion-LM [19] is the first to adapt diffusion models for handling text in a continuous space, achieving direct integration of continuous noise into word embeddings through

embedding and rounding procedures. Subsequent efforts have concentrated on utilizing continuous text diffusion models for sequence-to-sequence tasks. One such approach, DiffuSeq [36], divides the input into two segments, using one as a conditioning element while introducing noise to perturb the other. Moreover, SeqDiffuSeq [37] and DINOISER [38] have incorporated an encoder-decoder architecture into diffusion models via cross-attention mechanisms. GENIE [16] has demonstrated significant performance improvements by leveraging pre-training methodologies on extensive text corpora, a common practice in natural language processing. Furthermore, auto-regressive diffusion [17] employs a multi-level diffusion strategy that addresses sequential dependencies in text generation by involving both sentence-level and token-level diffusion. Latent Diffusion model [39], [40] utilize the fundamental architecture of language models, known as the encoder-decoder, to implement a diffusion process on the hidden representation generated by the encoder before it is fed into the decoder. However, despite these advancements, the summaries generated by text diffusion models still exhibit factual inconsistencies. Abstract summarizers based on diffusion models have not yet been developed with the objective of addressing this issue of inconsistency with sources. This paper introduces a factuality-guided module designed to reduce factual inconsistencies in diffusion-based summarizers. This module does not require retraining of the diffusion summarizer and integrates seamlessly with the diffusion process, with the aim of increasing factual consistency while maintaining the diversity of the generated summaries.

III. PRELIMINARIES

A. DIFFUSION MODEL

The diffusion model [41] is a probabilistic model which models data using two processes: a forward noising process and a reverse denoising process. Given data sampled from a distribution, the data is encoded as a latent representation $\mathbf{z}_0 \in \mathbb{R}^d$. Then, the forward process gradually corrupts $\mathbf{z}_0$ until it becomes a Gaussian distribution $\mathbf{z}_T \sim \mathcal{N}(0, I)$ at diffusion step T. Each step of the forward process is defined as:

$$q(\mathbf{z}_t|\mathbf{z}_{t-1}) = \mathcal{N}(\mathbf{z}_t; \sqrt{1-\beta_t}\mathbf{z}_{t-1}, \beta_t\mathbf{I}), \tag{1}$$

where β_t represents the level of noise introduced at timestep t, and $\mathbf{z}_t$ denotes the latent representation at timestep t.

The reverse process starts from the standard Gaussian noise $\mathbf{z}_T \sim \mathcal{N}(0, I)$ and then progressively denoises the noise to reconstruct $\mathbf{z}_0$. At each denoising step, denoising model $p_\theta(\mathbf{z}_{t-1}|\mathbf{z}_t)$, a transition from $\mathbf{z}_t$ to $\mathbf{z}_{t-1}$, is parameterized:

$$p_\theta(\mathbf{z}_{t-1}|\mathbf{z}_t) = \mathcal{N}(\mathbf{z}_{t-1}; \mu_\theta(\mathbf{z}_t, t), \Sigma_\theta(\mathbf{z}_t, t)),$$

where μ_θ and Σ_θ represent the mean and variance that those are learned from a neural network, respectively. Following Li et al. [19], it is possible to predefine the variance without the learning.

The learning objective for the diffusion model is derived from the variational lower bound of the negative log-likelihood of $\mathbf{z}_0$. Ho et al. [41] simplifies the objective to:

$$\mathcal{L}(\mathbf{z}_0 = \sum_{t=1}^{T} \mathbb{E}_{q(\mathbf{z}_t|\mathbf{z}_0)}||\mu_\theta(\mathbf{z}_t, t) - \tilde{\mu}(\mathbf{z}_t, \mathbf{z}_0)||^2,$$

where $\tilde{\mu}_t(\mathbf{z}_t, \mathbf{z}_0)$ denotes the mean of forward process posteriors $q(\mathbf{z}_{t-1}|\mathbf{z}_t, \mathbf{z}_0)$. That is, the diffusion model is trained to predict the forward process posteriors mean.

B. DIFFUSION MODEL FOR ABSTRACTIVE SUMMARIZATION

Given a source text $\mathbf{x} = (x_1, \ldots, x_m)$ consisting of m tokens, the goal of abstractive summarization is to generate a concise summary text $\mathbf{y} = (y_1, \ldots, y_n)$, such that $n \leq m$. Here, the summary should capture essential information from the source text. Diffusion-based abstractive summarizer models [16], [17], denoted as $p_\theta(\mathbf{y}|\mathbf{x})$, aims to generate the summary $\mathbf{y}$ conditioned on the source $\mathbf{x}$ through two processes. In the forward noising process, the diffusion-based summarizer gradually corrupts the summary in a continuous space by adding random noise. In the reverse process, it reconstructs the summary from random Gaussian noise using multiple denoising steps, each conditioned on the source text. During inference, this reverse process begins with sampling noise from a Gaussian distribution and iteratively denoises it with reference to the source text to generate the final summary.

As the diffusion process generally proceeds in the continuous space, it is necessary to map the discrete token sequence into this space. To this end, the diffusion-based summarization adopts an embedding function, which maps discrete $\mathbf{y}$ into a latent representation $\mathbf{z}_0$. This is represented as:

$$q_\theta(\mathbf{z}_0|\mathbf{y}) = \mathcal{N}(\mathbf{z}_0; \textsc{Emb}(\mathbf{y}), \beta\mathbf{I}).$$

Here, $\textsc{Emb}(\cdot)$ denotes the embedding function for the text sequence which is defined as

$$\textsc{Emb}(\mathbf{y}) = [Emb(y_1), \ldots, Emb(y_n)] \in \mathbb{R}^{n\times d}.$$

Each token in the summary text is embedded using the token embedding function *Emb*, and the resulting embedded tokens are then concatenated. Then, the forward process incrementally adds Gaussian noise to $\mathbf{z}_0$ until it becomes a standard Gaussian noise as outlined in Equation (1).

In the reverse process, the summary text $\mathbf{y}$ is reconstructed through denoising from sampled Gaussian noise, referring to the source text $\mathbf{x}$. For this, the encoder-decoder Transformer architecture [42] is adopted as the denoising model. At each denoising step, the decoder denoises $\mathbf{z}_t \in \mathbb{R}^{n\times d}$ based on the cross-attention with the encoded representations of $\mathbf{x}$ using the encoder. It is parameterized as:

$$p_\theta(\mathbf{z}_{t-1}|\mathbf{z}_t, \mathbf{x}) = \mathcal{N}(\mathbf{z}_{t-1}; \mu_\theta(\mathbf{z}_t, t, \mathbf{x}), \Sigma_\theta(\mathbf{z}_t, t, \mathbf{x})),$$

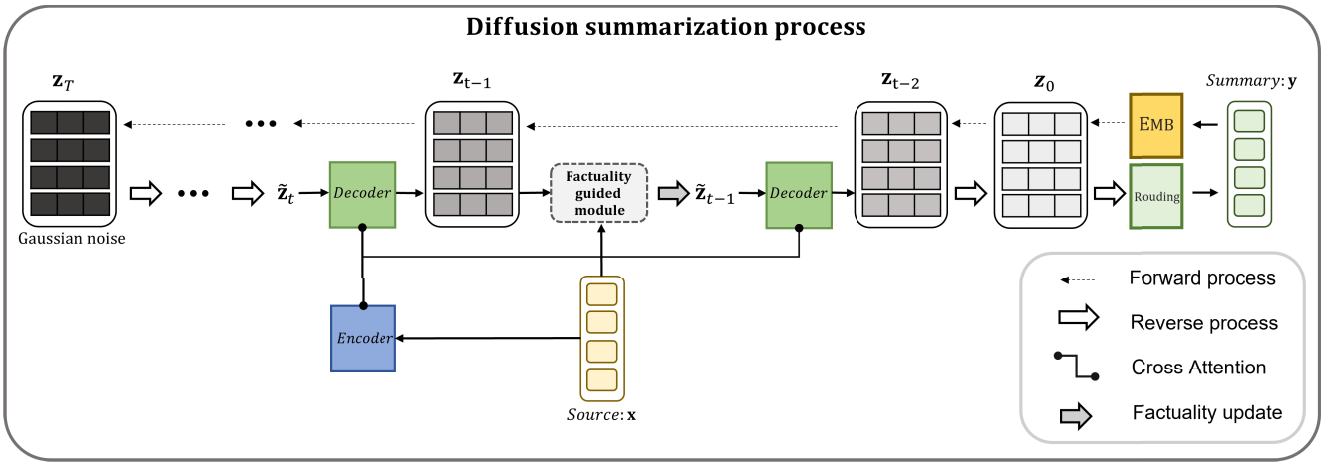

FIGURE 1. Overview of factuality guided diffusion-based abstractive summarization.

where the mean μ_θ and variance Σ_θ are predicted by a encoder-decoder Transformer network parameterized by θ. The reverse process is performed to generate a summary up to $\mathbf{z}_0$. Specifically, this involves generating a summary by iteratively denoising Gaussian noise $\mathbf{z}_t$ through T steps. Following this, to map vectors in the embedding space back to words, a discrete token is generated through a trainable rounding step, detailed as follows:

$$p_\theta(\mathbf{y}|\mathbf{z}_0) = \Pi_{i=1}^{n} p_\theta(y_i|z_0^i), \tag{2}$$

where $z_0^i \in \mathbb{R}^d$ is i-th vector of $\mathbf{z}_0$. That is, $\mathbf{z}_0 = \oplus_{i=1}^{n} z_0^i$, where $\oplus$ is a concatenate operator. $p_\theta(y_i|z_0^i)$ is a softmax distribution with the linear layer mapping to d-dimension to a vocabulary dimension. z_0^i is mapped by a rounding step to the most corresponding word y_i in the vocabulary.

Several variants of diffusion models have been developed by modifying the forward and reverse processes. One such variant, auto-regressive diffusion [17], employs a multi-level diffusion strategy that accounts for sequential dependencies in text generation. Specifically, the forward process of auto-regressive diffusion introduces noise more rapidly to the tokens on the right side of a sentence, transitioning from token embedding to Gaussian noise, while the tokens on the left side experience a slower noise addition. On the other hand, during the reverse process, the Gaussian noise at the positions on the left side is removed more quickly. This asymmetry has the advantage of enabling the right-side tokens to more effectively utilize the contextual information of the left-side tokens.

IV. FACTUALITY GUIDED DIFFUSION-BASED ABSTRACTIVE SUMMARIZATION

A. GUIDING THE SUMMARIZATION PROCESS

The goal of factuality-guided text summarization is to generate $\mathbf{y}$ from a conditional distribution $p(\mathbf{y}|\mathbf{x}, \mathbf{c})$, given the source text $\mathbf{x}$ and a factuality condition $\mathbf{c}$. According to the study of [19], generating $\mathbf{y}$ with a condition $\mathbf{c}$ can be formulated as a conditional diffusion model, which is

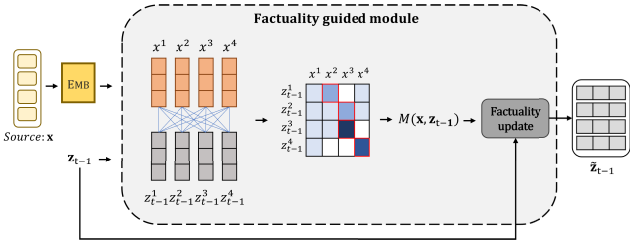

FIGURE 2. Token matching-based factuality guided module.

equivalent to decoding from the posterior $p(\mathbf{y}|\mathbf{x}, \mathbf{c}) = p_\theta(\mathbf{y}|\mathbf{z}_0) \prod_{t=1}^{T} p(\mathbf{z}_{t-1}|\mathbf{z}_t, \mathbf{x}, \mathbf{c})$. At each denoising step, the term $p(\mathbf{z}_{t-1}|\mathbf{z}_t, \mathbf{x}, \mathbf{c})$ is decomposed into a sequence of conditional problems using Bayes' formulation as follows:

$$p(\mathbf{z}_{t-1}|\mathbf{z}_t, \mathbf{x}, \mathbf{c}) \propto p_\theta(\mathbf{z}_{t-1}|\mathbf{z}_t, \mathbf{x}) p(\mathbf{c}|\mathbf{z}_{t-1}, \mathbf{z}_t, \mathbf{x}). \tag{3}$$

In this equation, $p_\theta(\mathbf{z}_{t-1}|\mathbf{z}_t, \mathbf{x})$ is estimated using a diffusion-based abstractive summarization model and $p(\mathbf{c}|\mathbf{z}_{t-1}, \mathbf{z}_t, \mathbf{x})$ is estimated by the factuality-guided module which will be described in the next section. Based on the decomposition, this paper adopts plug-and-play approaches [18] that keep the pre-trained diffusion model frozen and introduce a factuality-guided module to update the generated latent representation $\mathbf{z}_t$ during the reverse process.

Figure 1 shows how the proposed factuality-guided module works with the trained diffusion-based summarizer. At each step t, the decoder first samples the denoised representation $\mathbf{z}_{t-1}$ as described in Section III-B. Then, the proposed factuality-guided module calculates the factuality between the source text $\mathbf{x}$ and $\mathbf{z}_{t-1}$ and perform the guided correction to get representation $\tilde{\mathbf{z}}_{t-1}$. The updated $\tilde{\mathbf{z}}_{t-1}$ is used as input for the subsequent denoising step. As a result of the factuality guided module, the intermediate latent representation continually considers the factual consistency with the source text which results in a final summary that is factually consistent.

B. THE FACTUALITY GUIDED MODULE

The factuality-guided module estimates $p(\mathbf{c}|\mathbf{z}_{t-1}, \mathbf{z}_t, \mathbf{x})$ in Equation (3) and applies it to update the representation $\mathbf{z}_{t-1}$ sampled from $p_\theta(\mathbf{z}_{t-1}|\mathbf{z}_t, \mathbf{x})$ by the trained diffusion model. By the conditional independence assumption, $p(\mathbf{c}|\mathbf{z}_{t-1}, \mathbf{z}_t, \mathbf{x})$ is simplified to $p(\mathbf{c}|\mathbf{z}_{t-1}, \mathbf{x})$ and this paper defines it as follows:

$$p(\mathbf{c}|\mathbf{z}_{t-1}, \mathbf{x}) = \frac{1}{Z}(\mathbf{c} - M(\mathbf{x}, \mathbf{z}_{t-1})), \tag{4}$$

where Z is a normalizing factor and M quantifies the factuality between the source text $\mathbf{x}$ and the latent representation $\mathbf{z}_{t-1}$. If $\mathbf{c}$ is set to the positive constant, the value of M increases as $\mathbf{z}_{t-1}$ becomes more factually consistent with $\mathbf{x}$. Therefore, the fact-based module updates $\mathbf{z}_{t-1}$ at each reverse step to increase the value of M.

Since M function in Equation (4) takes $\mathbf{z}_{t-1}$ as input, it is computed in continuous space. To do this, this paper designs the function M as a token-level contextual matching between the source text and the latent representation. Figure 2 shows the process of token-level contextual matching of the proposed factuality guided module. Note that during inference, the reverse process begins with the sampling noise $\mathbf{z}_T \in \mathbb{R}^{\hat{n}\times d}$ from a Gaussian distribution, where $\hat{n}$ represents the length of the summary to be generated. This noise is then progressively denoised to reconstruct $\mathbf{z}_0$. Each latent representation $\mathbf{z}_*$ is represented as a sequence of d-dimensional vectors z_*^i because each vector z_*^i corresponds to the i-th token of the final summary as determined through the rounding step in Equation (2). This implies that the similarity between the source text and the latent representation can be quantified at the token-level. Moreover, this token-level similarity has a strong correlation with human judgment of factuality in document summarization [43], [44].

Formally, each token in the source text $\mathbf{x}$ is matched with a token in the latent representation $\mathbf{z}_{t-1}$ to calculate the similarity. Then, a greedy matching is executed to maximize the similarity, where each token is matched to the most similar token in other representation. That is, the M function is thus defined as:

$$R(\mathbf{x}, \mathbf{z}_{t-1}) = \frac{1}{m}\sum_{x_i\in\mathbf{x}} w(x_i) \max_{z_{t-1}^j \in \mathbf{z}_{t-1}} Emb(x_i)^\top z_{t-1}^j,$$

$$P(\mathbf{x}, \mathbf{z}_{t-1}) = \frac{1}{\hat{n}}\sum_{z_{t-1}^j\in\mathbf{z}_{t-1}} \max_{x_i\in\mathbf{x}} Emb(x_i)^\top z_{t-1}^j,$$

$$M(\mathbf{x}, \mathbf{z}_{t-1}) = R(\mathbf{x}, \mathbf{z}_{t-1}) \cdot P(\mathbf{x}, \mathbf{z}_{t-1}).$$

Here, $Emb(\cdot)$ is the embedding function from the diffusion-based summarizer, z_{t-1}^j is j-th vector of $\mathbf{z}_{t-1}$ and R and P functions can be interpreted as recall and precision, respectively. $w(\cdot)$ in R denotes importance weighting of token to consider the significance of tokens with infrequent occurrences when computing the similarity. According to the previous studies on similarity calculation [20], [45], [46], rare tokens can be more representative for similarity than common words. This paper adopts a normalized inverse document frequency computed from the training corpus as importance weights [20].

Once the factuality score $p(\mathbf{c}|\mathbf{z}_{t-1}, \mathbf{z}_t, \mathbf{x})$ is estimated, the factuality-guided denoising process is performed according to the score-based diffusion model [47]. Using Equation (3), the score function $\nabla_{\mathbf{z}_{t-1}} \log p(\mathbf{z}_{t-1}|\mathbf{z}_t, \mathbf{x}, \mathbf{c})$ for the t-th denoising step can be decomposed into two terms, such as

$$\nabla_{\mathbf{z}_{t-1}} \log p(\mathbf{z}_{t-1}|\mathbf{z}_t, \mathbf{x}, \mathbf{c}) = \nabla_{\mathbf{z}_{t-1}} \log p_\theta(\mathbf{z}_{t-1}|\mathbf{z}_t, \mathbf{x}) + \nabla_{\mathbf{z}_{t-1}} \log p(\mathbf{c}|\mathbf{z}_{t-1}, \mathbf{x}). \tag{5}$$

In this equation, the first term is estimated using a diffusion-based abstractive summarization model. The second term acts as a correction gradient [47], [48] that directs $\mathbf{z}_{t-1}$ toward a hyperplane in the latent space, where all latent representations align with the given condition $\mathbf{c}$ and the source text $\mathbf{x}$.

Based on the decomposition, we first sample $p_\theta(\mathbf{z}_{t-1}|\mathbf{z}_t, \mathbf{x})$ in Equation (3) using a denoising step:

$$\mathbf{z}_{t-1} = (1 + \frac{1}{2}\beta_t)\mathbf{z}_t + \beta_t \nabla_{\mathbf{z}_t} \log p(\mathbf{z}_t, \mathbf{x}) + \sqrt{\beta_t}\epsilon$$

where $\epsilon \sim \mathcal{N}(0, I)$ is randomly sampled Gaussian noise and $\nabla_{\mathbf{z}_t} \log p(\mathbf{z}_t, \mathbf{x})$ is the score function estimated by the decoder in the diffusion summarization model. Then, the factuality-guided correction is performed by the sampling formula:

$$\tilde{\mathbf{z}}_{t-1} = (1 + \frac{1}{2}\beta_t)\mathbf{z}_{t-1} + \beta_t \nabla_{\mathbf{z}_{t-1}} \log p(\mathbf{c}|\mathbf{z}_{t-1}, \mathbf{x}) + \sqrt{\beta_t}\epsilon$$

where the new $\tilde{\mathbf{z}}_{t-1}$ then serves as the input for the next time step. This process is repeated until $t = 0$, ensuring the factuality between $\mathbf{x}$ and $\mathbf{z}_*$ is maintained. The final summary is generated from given $\tilde{\mathbf{z}}_0$ via the rounding step described in Equation (2).

TABLE 2. A simple statistics on the datasets used for the experiments.

Dataset	Train	Valid	Test	Avg. summary length
CNN/DM	286,817	13,368	11,487	296
XSUM	204,045	11,332	11,334	125
Gigaword	3,803,957	189,651	1,951	51

V. EXPERIMENTS

A. EXPERIMENTAL SETTINGS

The proposed summarizer is evaluated with three widely-used benchmark datasets: CNN/DM [49], XSUM [50], and Gigaword [51]. These datasets have primarily been employed in recent summary research, including diffusion-based summarizations [16], [17], [52]. For example, in the case of CNN/DM, the reference summaries, authored by humans, are used to assess the generated summaries. The CNN/DM dataset consists of over 300,000 news articles, written by journalists from CNN and the Daily Mail. The XSUM dataset includes a diverse range of BBC articles from 2010 to 2017, covering domains such as news, politics, and sports. The Gigaword dataset, designed for predicting

relationships between articles and their headlines, is adapted for summarization by considering articles as original texts and their headlines as summary references. Table 2 presents the statistics of these datasets. This paper follows the official data splits for evaluating abstractive summarization models, given their standard usage in the field.

The proposed method is first compared against two baselines in Section V-B1: an autoregressive (AR) based summarizer and a diffusion-based summarizer. The AR-based summarizer, a pre-trained transformer-based seq2seq model, is fine-tuned for each dataset. The diffusion-based summarizer generates summaries by denoising the random sample using a non-autoregressive transformer. Specifically, the BART [6] model is used as the AR-based summarizer and the AR-diffusion [17] and GENIE [16] are adopted as diffusion-based models. Next, in Section V-B2, we compared the proposed method with large language models (LLMs) that have recently demonstrated remarkable performance. Specifically, we conducted comparisons using a zero-shot setup and LoRA fine-tuning for two LLM models: Llama2 7b and Mistral 7b. Finally, in Section V-B3, a qualitative evaluation was performed using LLMs, which are increasingly popular for assessing the quality of machine-generated texts.

This paper follows the generation protocol proposed by [17]. That is, for each source text, diffusion-based summarizers randomly sample 50 Gaussian noises and generate 50 summaries from the sampled noises. For AR-based summarizer and LLMs, decoding methods such as beam search, top-k sampling, and top-p sampling are used to generate 50 summaries from the same source text. To select the best summary from the generated summaries, this paper adopts Minimum Bayes Risk [53] decoding. That is, given the generated summaries $\mathcal{Y}$, it selects the summary $\hat{\mathbf{y}}$ that achieves the minimum expected risk under a loss function $\mathcal{L}$ such that

$$\hat{\mathbf{y}} = \arg\max_{\mathbf{y}\in\mathcal{Y}} \sum_{\mathbf{y}'\in\mathcal{Y}} \frac{1}{|\mathcal{Y}|}\mathcal{L}(\mathbf{y}, \mathbf{y}').$$

Here, $\mathcal{L}$ is used as the ROUGE-1 score. Given that a higher ROUGE-1 score indicates better performance, the formula is modified to argmax.

This paper evaluates the proposed model and its baselines across three dimensions: lexical overlap, factual consistency, and diversity. Lexical overlap is measured by comparing the generated summaries with reference summaries using the ROUGE [54]. Factual consistency is evaluated using two automatic metrics: BERTScore [20] and QFE [55]. Lastly, for diversity, the models generate multiple summary sentences, and the variety among these sentences is quantified using the SELF-BLEU [56]. SELF-BLEU is based on the BLEU score, which assesses the similarity between two sentences. It can also evaluate how closely one generated summary resembles the rest among multiple summary sentences. By treating one sentence as the hypothesis and the others as references, we can calculate the BLEU score for each generated summary and define the average BLEU score as the SELF-BLEU of the multiple summary sentences. The lower the SELF-BLEU score, the more diverse the generated summaries are. It is important to note that optimizing for factual consistency does not necessarily guarantee diversity. For instance, some models might generate diverse summaries by making minor alterations to a summary that already has high factual consistency. Ideally, both factual consistency and diversity should be maximized, while acknowledging any trade-offs between them. Therefore, this paper introduces F1_{BS}, the harmonic mean of factual consistency and diversity, analogous to how recall and precision are combined in F1. It is defined as follows:

$$\text{F1}_{\text{BS}} = 2 \times \frac{\text{BERTScore} \times (100 - \text{SELF-BLEU})}{(\text{BERTScore} + (100 - \text{SELF-BLEU}))}.$$

The reason for subtracting SELF-BLEU from 100 is because the range of the SELF-BLEU score is from 0 to 100 and the lower is preferable.

The proposed method utilizes the AR-diffusion summarizer, a diffusion-based summarizer that is pretrained on each abstractive summarization dataset. The underlying architecture is an encoder-decoder transformer model, which includes a 6-layer encoder and a 6-layer cross-attention decoder as the denoising architecture. Each layer contains 8 attention heads and a hidden dimension of 1,024. The diffusion process is configured with an embedding dimension of 128, a square root noise schedule β_t, and 2,000 diffusion steps (T). This number of steps are matched in the AR-diffusion model. For evaluation, all diffusion-based summarizers utilize the tokenizer and vocabulary from bert-base-uncased. In Equation (4), the factuality condition $\mathbf{c}$ is set to 1 to ensure the most faithful representation of the source text. All experiments are conducted using four GTX-3090 GPUs.

B. EXPERIMENTAL RESULTS

1) PERFORMANCE OF SUMMARIZATION

Table 3 presents performances of the proposed summarizer and baselines across the CNN/DM, XSum, and Gigaword datasets. First, the proposed summarizer outperforms other baselines in terms of factual consistency. The most significant margins are observed on the XSUM and Gigaword datasets. The proposed summarizer achieves a BERTScore of 63.08 on the XSUM dataset, which is 3.26 points higher than that of the BART model using beam search. Similarly, on the Gigaword dataset, it achieves a score of 72.85, outperforming the BART model by 1.4 points. The summaries from the XSUM and Gigaword datasets are highly abstract and are written in a form that differs significantly from the source text. The proposed summarizer, equipped with a factuality-guided module, consistently refines intermediate summaries to enhance factual consistent with the source text. As a result, it generates a final summary that is both coherent and factually consistent.

Moreover, the proposed summarizer achieves higher ROUGE scores compared to the baselines across XSum, and

TABLE 3. Performance measured with BERTScore, QFE and ROUGE on CNN/DM, XSUM, and Gigaword datasets.

Dataset	Model	Sampling	BERTScore (↑)	QFE (↑)	SELF-BLEU (↓)	$F1_{BS}$ (↑)	ROUGE 1	ROUGE 2	ROUGE L
CNN/DM	BART	beam	86.73	2.7953	95.41	8.72	**40.74**	**17.22**	37.45
		top-p	84.91	2.5765	83.87	27.11	37.52	15.77	34.26
		top-k	84.27	2.5193	84.01	26.88	36.83	14.95	33.67
	GENIE	Diffusion	82.88	2.4558	**37.82**	71.05	34.4	12.8	32.2
	AR-diffusion		84.92	2.6493	38.12	71.59	39.6	16.3	37.2
	Proposed Method		**87.14**	**2.8272**	38.23	**72.29**	40.58	17.06	**37.52**
XSUM	BART	beam	59.82	0.8107	90.72	11.51	31.81	10.47	25.03
		top-p	58.26	0.7819	76.24	24.96	29.37	7.96	21.62
		top-k	57.94	0.7386	76.62	24.51	28.87	7.58	21.15
	GENIE	Diffusion	56.84	0.6983	**29.51**	50.84	29.2	8.1	21.5
	AR-diffusion		57.58	0.7136	31.52	50.68	31.6	10.1	24.7
	Proposed Method		**63.08**	**0.9214**	31.83	**55.38**	**32.18**	**10.97**	**25.08**
Gigaword	BART	beam	71.45	1.0958	89.12	15.86	31.15	13.49	28.13
		top-p	70.84	0.9751	73.37	33.16	30.15	12.91	27.48
		top-k	70.28	0.9583	75.73	30.62	30.87	12.42	27.08
	GENIE	Diffusion	69.58	0.9087	**41.19**	56.07	29.78	11.32	26.92
	AR-diffusion		70.91	1.095	41.52	56.95	30.18	12.95	27.53
	Proposed Method		**72.85**	**1.1877**	41.81	**58.34**	**31.39**	**13.73**	**28.53**

TABLE 4. Performances of LLM-based summarization on CNN/DM and and XSUM datasets.

Dataset	Model	Sampling	BERTScore (↑)	QFE (↑)	SELF-BLEU (↓)	$F1_{BS}$ (↑)	ROUGE 1	ROUGE 2	ROUGE L
CNN/DM	Llama2 7b (Zero-shot)	beam	75.13	2.4257	87.45	18.92	26.38	10.10	24.84
		top-p	71.07	2.1242	55.48	48.06	22.83	5.94	20.89
		top-k	70.87	2.1141	56.67	47.07	22.74	5.56	21.12
	Llama2 7b (Fine-tuned)	beam	85.21	2.5848	88.45	19.94	31.42	12.74	29.45
		top-p	82.02	2.4948	80.35	30.18	25.01	7.47	22.42
		top-k	82.94	2.4750	80.46	30.38	25.32	7.82	22.13
	Mistral 7b (Zero-shot)	beam	85.23	2.9216	88.49	19.89	39.47	16.51	35.41
		top-p	81.76	2.4851	78.41	32.47	36.95	12.67	32.02
		top-k	82.14	2.5258	78.07	33.03	37.03	13.17	32.47
	Mistral 7b (Fine-tuned)	beam	86.43	**2.9738**	89.90	17.95	40.31	**17.52**	37.28
		top-p	82.67	2.7653	86.86	21.67	38.94	14.63	34.73
		top-k	82.38	2.7328	85.99	22.82	38.89	14.84	34.83
	Proposed Method	Diffusion	**87.14**	2.8272	**38.23**	**72.29**	**40.58**	17.06	**37.52**
XSUM	Llama2 7b (Zero-shot)	beam	78.42	0.8841	90.79	14.99	31.32	14.17	26.13
		top-p	73.21	0.8028	67.52	39.76	20.56	4.75	16.53
		top-k	71.93	0.7913	66.88	49.62	21.12	5.03	17.12
	Llama2 7b (Fine-tuned)	beam	**83.04**	0.9135	95.51	8.14	31.52	**14.05**	**26.21**
		top-p	79.53	0.8975	73.53	37.06	22.13	5.04	17.19
		top-k	78.84	0.8841	75.23	34.90	21.68	6.12	17.25
	Mistral 7b (Zero-shot)	beam	80.38	0.8931	90.51	13.82	27.91	8.63	22.54
		top-p	66.19	0.8451	79.31	25.40	24.76	7.66	18.92
		top-k	66.41	0.8641	80.03	24.76	24.47	7.87	18.26
	Mistral 7b (Fine-tuned)	beam	82.51	0.9073	90.74	15.85	30.84	13.78	25.08
		top-p	74.32	0.8625	79.14	28.71	27.82	10.67	19.72
		top-k	74.56	0.8751	78.94	29.02	28.12	9.61	20.21
	Proposed Method	Diffusion	63.08	**0.9214**	**31.83**	**65.52**	**32.18**	10.97	25.08

Gigaword datasets. It records the highest ROUGE-1 scores of 32.18, and 31.39 for the XSum, and Gigaword datasets, respectively. The proposed summarizer has 0.16 points lower ROUGE-1 than the BART model using beam search in CNN/DM data. This is a very small difference, and rather, it shows a small but superiority in ROUGE-L. It is noteworthy that the proposed model achieves higher ROUGE scores than the AR-diffusion model, which indicates that the proposed model generates summaries that are similar to the reference summaries.

In terms of diversity, the proposed summarizer shows superior SELF-BLEU performance compared to the BART

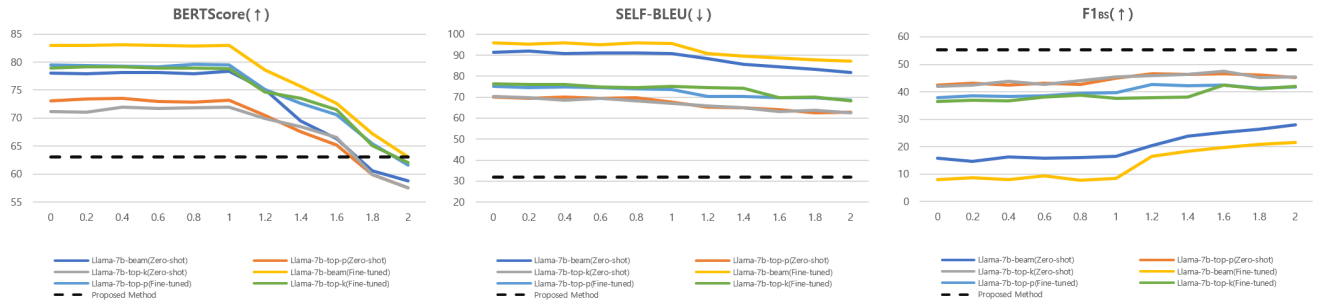

FIGURE 3. Performance of LLMs in terms of diversity and factuality at different temperatures.

TABLE 5. Qualitative evaluation using large language model on XSUM dataset.

Model	BART		GENIE	AR-diffusion	Proposed Method
Sampling	top-p	top-k	Diffusion		
Average summary score	2.56	2.51	2.54	2.54	2.65
Average high-quality summary	6.11	6.01	6.03	6.10	6.18

TABLE 6. Performance of LLMs based on the number of parameters.

Model	BERTScore	SELF-BLEU	$F1_{BS}$
Llama2-7b-beam	78.42	90.79	16.48
Llama2-7b-top-p	73.21	67.52	44.99
Llama2-7b-top-k	71.93	67.52	45.35
Llama2-13b-beam	79.25	91.25	15.76
Llama2-13b-top-p	74.52	69.51	43.27
Llama2-13b-top-k	73.94	67.32	45.33
Llama2-30b-beam	80.27	88.56	20.03
Llama2-30b-top-p	75.69	68.54	44.45
Llama2-30b-top-k	76.02	67.51	45.52
Proposed Method	63.08	31.83	65.52

TABLE 7. The prompt used for Qualitative evaluation.

Prompt Input > You are a summary evaluator.

Police have confirmed paint was thrown over doors and windows at Crumlin Orange Hall some time between 2300 BST on Monday and 1000 BST on Tuesday. SDLP South Antrim MLA Thomas Burns condemned those behind the attack which he said was "pointless". "They can only harm and destroy, they can only cause pain and hardship and useless expense to taxpayers and ratepayers," he said.

If the following sentences as summary of the above article, please assign an overall score. Scores range from 1 to 3, 1 represents bad, 2 represents neutral, 3 represents good. The output format is 'Score: 1'.

<Generated summary>

Score:

Output > 3

summarizer. However, it shows slightly lower SELF-BLEU scores than the diffusion-based baselines (GENINE and AR-diffusion). This disparity is not significant and might be attributable to the increased factual consistency. In terms of $F1_{BS}$, the proposed model achieves the best performance across all three datasets. From these results, this paper validates that the proposed summarizer is capable of generating summaries that are both diverse and factually consistent.

2) A COMPARISON BETWEEN THE PROPOSED METHOD AND LLMS

Table 4 presents the performance of both the proposed models and the LLM-based summarizers on the CNN/DM and XSUM datasets. Similar to the AR based summarizer, all the LLMs have high factual evaluation scores, BERTScore and QFE, in the beam-search sampling. Using top-k and top-p sampling as a way to increase diversity produces more diverse summaries, which improves the SELF-BLUE score, but at the expense of factual consistency. This phenomenon is observed in both zero-shot and fine-tuned models. In the fine-tuned models, factual scores are higher than in the zero-shot models, but diversity is reduced due to the specialization for the dataset.

On CNN/DM data, summaries from the proposed models demonstrate similar or higher factual consistency compared to those from the LLM models. Specifically, the proposed method has a BERTScore of 0.71 higher and a QFE of about 0.14 lower than Mistral 7b (fine-tuned) with beam sampling, which has the highest factual consistency score among the LLM models. Additionally, the proposed method significantly outperforms in diversity, with a SELF-BLEU score difference of 51.67 and a harmonic mean score $F1_{BS}$ difference of 54.34.

In the XSum dataset, unlike the CNN/DM dataset, the BERTScore of the proposed method is lower than that of the LLM models. This is because the LLM models have high BERTScore due to their tendency to copy the source. However, the proposed method achieves the highest QFE

TABLE 8. Examples of generated summaries in XSUM dataset. The underline in the summaries implies an inconsistent token span.

Source text	The US-born radical Yemeni cleric Anwar al-Awlaki is head of al-Qaeda in the Arabian Peninsula.Two brothers believed to be mid-ranking al-Qaeda officials died in a drone strike in south Yemen on Thursday, Yemeni officials said.The attack came just days after al-Qaeda chief Osama Bin Laden was killed in Pakistan by US Navy Seals.The Pentagon refused to comment on the reports that Anwar al-Awlaki was specifically targeted in Yemen.According to Yemen's defence ministry, the missile fired by the drone hit a car in in the province of Shabwa carrying two brothers, identified by Yemeni officials as Musa'id and Abdullah Mubarak. ...
BART with beam search	0. al-Qaeda officials has been killed in a us drone strike in yemen, us media say. 1. a top al-qaeda leader has been killed in a us drone strike in yemen, us media reports say.
BART with top-k	0. the us military officials say they aimed to kill the head of al-qaida's yemeni branch, reports from the wall street journal say. 1. the us military officials say to kill the head of al-qaeda in yemen in an air strike last week, the wall street journal say.
BART with top-p	0. the us military has allegedly carried out a drone strike on an al-qaeda leader in yemen. 1. the us military officials say their drone strike on an al-qaeda leader may have hit a family in yemen but failed to kill him.
AR-diffusion	0. a drone strike has killed head of al-qaeda in south yemen on thursday, officials say. 1. pentagon chose not to respond to reports suggesting that was targeted by yemen's defence ministry.
Proposed method	0. a drone strike has killed two brothers believed to hold mid-ranking in the yemen, on thursday. 1. pentagon refused to comment on the reports that US-born yemeni cleric anwar al-awlaki was targeted in yemen.
Source text	George Verrier was treated by officers who were called to an altercation involving about 20 people in Bromley in the early hours of Sunday.He did not want to be taken to hospital, said police, but was found unconscious several hours later in nearby Ferndale. He died in hospital.A 17-year-old arrested on suspicion of murder has been released on bail.The Met said it had voluntarily referred the actions of officers who went to the scene to the Independent Police Complaints Commission (IPCC) for assessment. A Metropolitan Police spokesman said: "At this early stage it appears the victim suffered a head injury as a result of the altercation."Inquiries continue to establish the full circumstances of the incident. George, a trainee electrician, was injured in Southborough Road at about 00:45 BST, while on his way home from the party in adjacent Blenheim Road. ...
BART with beam search	0. a 16-year-old boy has died after suffering a head injury at a party in south-west london. 1. a 19-year-old man has died after suffering a head injury at a house party in south-east london.
BART with top-k	0. police have referred the actions of officers at a party where a 16-year-old boy was injured to the police watchdog. 1. police are to be investigated after a 16-year-old boy died from a head injury he suffered at a party in south-west london.
BART with top-p	0. a man has died after being assaulted at a party in south london. 1. an 18-year-old man who died after suffering a hit to the head while on his way home from a house party in west london has been named over the loss of his life.
AR-diffusion	0. 17-year-old man has died after a fight at a party in south-west london. 1. a murder investigation is under way after a man died stabbed to fall at a party party.
Proposed method	0. a 17-year-old arrested on suspicion of murder after a fight at a party in southborough. 1. a murder investigation is under way after a man died suffered a head injury at a party.

score of 0.9214, indicating superior factual consistency compared to the LLM models. As evidenced by the SELF-BLEU scores, LLM models generally exhibit low diversity across all sampling methods, whereas the proposed method achieves a high diversity score. Consequently, the proposed method scores 5.76 points higher than Llama2 7b (zero-shot), which is the highest among LLMs in terms of $F1_{BS}$ score. These results across the two datasets suggest that the proposed method demonstrates superior performance in both diversity and factuality compared to the LLM-based models.

The LLM model can increase diversity by adjusting both the temperature and the sampling method. Therefore, we empirically demonstrate performance changes in response to temperature variations. Figure 3 presents a comparative experiment with the proposed model, exploring the relationship between factual consistency and diversity in LLM-generated summaries when altering the temperature. For LLM-generated summaries, factual consistency remains relatively stable when the temperature is between 0 and 1 but tends to decrease when it exceeds 1. However, it is observed that the diversity score does not improve significantly. Notably, there is a marked difference in diversity generation between the proposed method and the LLM-based method. Although the summary generated using zero-shot top-p and top-k sampling with the LLM model achieves the highest $F1_{BS}$ score, it still does not surpass the score of the proposed method.

We also compared the performance of the proposed method to that of LLMs as the number of parameters increased. Table 6 presents a comparative experiment examining factual consistency and diversity relative to the parameter size of the LLMs. Although increasing the size of the LLMs results in a slight improvement in BERTScore, the enhancement in diversity is not substantial. Consequently, the improvement in $F1_{BS}$ is not significant and remains lower than that achieved by the proposed method. This experiment suggests that merely increasing the number of parameters in LLMs does not necessarily result in better performance in terms of both factual consistency and diversity.

3) QUALITATIVE EVALUATION USING LARGE LANGUAGE MODEL

Abstractive summarization models may not be sufficiently evaluated using only automatic metrics [57] and often require qualitative evaluation as complements. Recently, large language models (LLMs) have gained popularity for evaluating the quality of machine-generated texts [58], [59], because human evaluation, though precise, requires

TABLE 9. Examples of generated summaries in XSUM dataset. The underline in the summaries implies an inconsistent token span.

Source text	Scores more are still missing in Teluk Wondama district after flash floods and landslides that were triggered when a river burst its banks. Hundreds of people have been injured and more than 5,000 are being housed in tents set up in emergency camps. The National Search and Rescue Agency and military have sent reinforcements to help deal with the disaster. The sudden flooding on Monday caught residents by surprise, officials said, contributing to the high death toll. The local airport has now reopened and roads have been largely cleared of debris, which is helping to speed up aid deliveries to the area, reports say. Navy warships had previously been used to carry tents, medical supplies and food to the disaster zone. Thousands of buildings, roads and bridges in five villages have been damaged or destroyed. The head of the the provincial relief agency, Dortheis Sawaki, said 104 bodies had been recovered, but many more people were missing and the death toll was expected to rise. ...
Llama2 7b with beam search (fine-tuned)	0. Flash flood and landslides in Teluk Wondama district have resulted in over 100 deaths, with hundreds injured. 1. Flash floods and landslides in Teluk Wondama district have caused extensive damage and resulted in over 100 deaths.
Llama2 7b with top-k (fine-tuned)	0. A major landslide has killed at least 90 people in Indonesia's Papua region. 1. fresh flooding has hit the Indonesian province of Papua, where more than 100 people died in a similar disaster last month.
Llama2 7b with top-p (fine-tuned)	0. At least 131 people have been killed by flooding and landslides in a remote part of the Indonesian province of Papua. 1. The death toll from a major storm that hit Indonesia's Papua province last week has risen to at least 139, local officials say.
Mistral 7b with beam search (fine-tuned)	0. Flash floods and landslides in Teluk Wondama district, Indonesia, have resulted in over 100 deaths, with hundreds injured and thousands displaced. The disaster was caused by a river bursting its banks, catching residents by surprise. 1. Flash floods and landslides in Teluk Wondama district, Indonesia, have resulted in over 100 deaths, with hundreds injured and thousands displaced. The disaster was caused by a river bursting its banks, which caught residents by surprise
Mistral 7b with top-k (fine-tuned)	0. Flash floods and landslides in Teluk Wondama district have resulted in the deaths of over 100 people, with thousands more missing. Over 5,000 people have been injured and are being housed in emergency camps. 1. Flash floods and landslides caused by a burst river in Teluk Wondama, Indonesia, have left over 100 people dead and thousands more missing. Hundreds of people were injured, and over 5,000 were housed in emergency camps.
Mistral 7b with top-p (fine-tuned)	0. Flash flood and landslides in Teluk Wondama district have resulted in over 100 deaths, with hundreds injured and thousands left homeless. Over 5,000 people are being housed in emergency camps. 1. Flash floods and landslides in Teluk Wondama district have resulted in the deaths of over 100 people, with thousands more missing. Over 5,000 people have been injured and are being housed in emergency camps.
Proposed method	0. a drone strike has killed two brothers believed to hold mid-ranking in the yemen, on thursday. 1. pentagon refused to comment on the reports that US-born yemeni cleric anwar al-awlaki was targeted in yemen.

specialized expertise and is both costly and challenging to reproduce consistently. Following on studies [60], [61] that demonstrate a close alignment of LLM-based evaluations with human assessments, this paper adopts a large language model to evaluate the quality of the generated summaries.

Considering our aim to generate summaries that are both diverse and factually consistent, directly comparing generated summaries to a single reference may not fully capture the quality of the summarization models. Therefore, instead of making direct comparisons between the summaries and a reference, this paper employs a LLM to provide summarization scores as a zero-shot setting. Specifically, for each source text and its corresponding generated summary, the LLM predict a score that represents the summary's quality. To facilitate this, this paper utilizes a specific prompt outlined in Table 7 with the `text-davinci-003` model. Note that each summary is assigned a score ranging from 1 to 3.

This paper follows the LLM-based evaluation framework proposed by [16], with the following specifics. Initially, it randomly samples 10% of the source text from the XSUM test data. For each source text, 10 summaries are generated for both the proposed model and the baselines. Subsequently, the LLM predicts the score for each summary. Performances are evaluated using two metrics: Average Summary Score and Average High-Quality Summary. The average summary score represents the mean score of all summaries, while the average high-quality summary refers to the average number of summaries receiving a score of 3.

Table 5 shows the scores of the proposed model and its baselines. The proposed summarizer shows high performance in both average summary score and average high-quality summary compared to other baselines. Results from average summary score demonstrate that the proposed summarizer consistently generates a significant number of high-quality summaries. Furthermore, scores from the average high-quality summary indicate that the majority of summaries generated from the proposed model are of good quality. These results imply that the proposed model is effective in generating summaries that are not only diverse but also of high quality.

4) CASE STUDY

Table 8 presents two sample summaries generated by the proposed summarizer alongside those from baseline models. Summaries generated using the BART model with beam search often demonstrate a lack of diversity, as evidenced by the repetition of similar words across different summaries. Efforts to enhance diversity in BART via top-k and top-p sampling methods have led to increased diversity. Nonetheless, these summaries sometimes include inconsistent tokens and exhibit diversity primarily in the form of minor word variations within a single topic. For instance, the first top-k summary mentions ''an al-Qaeda leader may have hit a family'', which is not present in the source text. A similar

TABLE 10. An example of failures and limitations.

Repetition	jordon ibe has broken into the first team this this season
Grammar error	Jose Mourinho's side extended their lead top of Barclays Premier League to seven points with a 2-1 win

issue is observed in the second summary from the top-p model, where ''south-west London'' are generated, also factually inconsistent with the source.

In contrast, summaries from diffusion-based summarizers show considerable diversity and employ different expressions for similar meanings, such as ''the us army'' and ''pentagon''. However, the AR-diffusion model often generates sentences with inconsistencies, like ''head of al-qaeda'' or ''a man died stabbed to fall'', neither of which align with the source text. On the other hand, the proposed summarizer consistently generates factually consistent summaries, such as ''two brothers'' and ''a man died suffered a head injury''. Since the proposed summarizer guides the reverse process toward increased factual alignment with the source, it generates summaries that are not only consistent with the source text but also high in diversity.

Table 9 provides a comparative overview of summaries generated by the LLM-based summarizer and our summarizer. For instance, the Llama2 7b model with beam sampling produced summaries with high factual consistency but limited diversity, often changing only a few words within the same topic. Although top-k and top-p sampling methods aimed for a wider variety of summaries, they inaccurately generated numbers such as ''90'' or ''139''. Mistral tended to generate longer texts than Llama2 but with lower diversity, focusing on similar topics. Despite top-p and top-k samplings offering higher factual consistency than Llama models, they sometimes resulted in incorrect summaries such as ''Over 5,000 people have been injured.'' On the other hand, the proposed model can generate summaries that not only covered a diverse range of topics and vocabulary, like ''Many people are still missing'' and ''104 bodies had been recovered,'' but also maintained high levels of factual accuracy.

5) FAILURE CASES AND LIMITATIONS

The diffusion-based summarizer introduced in this study incorporates a non-autoregressive decoder, a choice that inherently retains the token repetition and syntactical errors commonly associated with non-autoregressive models. Table 10 shows some examples of these failure cases. For instance, in the first example provided, there is a repetition of the tokens ''this'', and in the second example, the phrase ''at the top of Barclays Premier League'' lacks the grammatical correctness due to missing prepositions and articles. These issues stem from the inherent structure of token generation, where each token's creation is heavily influenced by its adjacent tokens. An analysis of 100 randomly chosen summary samples identified three instances of token repetition and five grammatical inaccuracies. Although these issues do not significantly impact the summaries semantic integrity or factual accuracy, addressing them is crucial for the enhancing the reliability of the summarization model.

VI. CONCLUSION

This paper introduces a factuality-guided module combined with a diffusion-based abstractive summarizer to generate summaries that are both diverse and factually consistent. The factuality module guides factuality at each denoising step, ensuring the diffusion-based summarizer reconstructs summaries that are both diverse and factually consistent. As the factuality-guided module integrates seamlessly in a plug-and-play fashion with the diffusion-based summarizer, it facilitates the generation of factually consistent summaries without the need for retraining the diffusion-based summarizer.

Experimental results from three benchmark datasets indicate that the proposed summarizer outperforms both pre-trained auto-regressive summarizers and contemporary state-of-the-art diffusion-based models. Moreover, the proposed summarizer has also demonstrated superior performance compared to summarizations generated by large language models. Additionally, qualitative evaluation using a large language model has confirmed the superiority of the proposed method. The experiments demonstrate that the factuality-guided diffusion-based abstractive summarizer is capable of generating summaries with high diversity and factual consistency.

The future work of this paper focuses on three main areas. First, we aim to measure the proposed factuality score not only at the token level but also at the sentence level. Measuring the factuality score at the sentence level will enhance the generation of summaries with high factuality by allowing comparisons between tokens within the context of their source sentences. Second, we plan to apply the proposed method to a consistency model [62] that increases generation speed by reducing the steps in the reverse process to one or a few steps. As there are currently few studies utilizing the consistency model in text generation, we will develop and propose a method that applies it to text generation to achieve high factuality. Finally, we intend to expand the proposed method to accommodate long source documents.

ETHICAL CONSIDERATIONS

The authors aim to generate diverse summaries while improving the factuality of the summarization model with respect to hallucinations, which is the main concern of the summary model. They find that summarization models tend to lack factuality when generating text with high diversity. Inaccurate summaries can be misleading to the user in terms of misinformation.

REFERENCES

[1] R. Gagiano, M. M.-H. Kim, X. J. Zhang, and J. Biggs, ''Robustness analysis of Grover for machine-generated news detection,'' in *Proc. 19th Annu. Workshop Australas. Lang. Technol. Assoc.* Australasian Language Technology Association, 2021, pp. 119–127.

[2] A. Mosallanezhad, K. Shu, and H. Liu, "Generating topic-preserving synthetic news," in *Proc. IEEE Int. Conf. Big Data*, Dec. 2021, pp. 490–499.
[3] W. Yu, C. Zhu, Z. Li, Z. Hu, Q. Wang, H. Ji, and M. Jiang, "A survey of knowledge-enhanced text generation," *ACM Comput. Surv.*, vol. 54, no. 11s, pp. 1–38, Nov. 2022.
[4] D. Liu, J. Li, M.-H. Yu, Z. Huang, G. Liu, D. Zhao, and R. Yan, "A character-centric neural model for automated story generation," in *Proc. AAAI Conf. Artif. Intell.*, Apr. 2020, vol. 34, no. 2, pp. 1725–1732.
[5] Y. Liu and M. Lapata, "Text summarization with pretrained encoders," in *Proc. Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process.*, 2019, pp. 3730–3740.
[6] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 7871–7880.
[7] D. Pernes, A. Mendes, and A. F. T. Martins, "Improving abstractive summarization with energy-based re-ranking," in *Proc. 2nd Workshop Natural Lang. Gener., Eval., Metrics (GEM)*. Abu Dhabi, UAE: Association for Computational Linguistics, 2022, pp. 1–17.
[8] D. Wan, M. Liu, K. McKeown, M. Dreyer, and M. Bansal, "Faithfulness-aware decoding strategies for abstractive summarization," in *Proc. 17th Conf. Eur. Chapter Assoc. Comput. Linguistics*. Dubrovnik, Croatia: Association for Computational Linguistics, 2023, pp. 2864–2880.
[9] Z. Cao, F. Wei, W. Li, and S. Li, "Faithful to the original: Fact-aware neural abstractive summarization," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018, pp. 4784–4791.
[10] T. Falke, L. F. R. Ribeiro, P. A. Utama, I. Dagan, and I. Gurevych, "Ranking generated summaries by correctness: An interesting but challenging application for natural language inference," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 2214–2220.
[11] C. Zhu, W. Hinthorn, R. Xu, Q. Zeng, M. Zeng, X. Huang, and M. Jiang, "Enhancing factual consistency of abstractive summarization," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, 2021, pp. 718–733.
[12] F. Nan, C. N. D. Santos, H. Zhu, P. Ng, K. McKeown, R. Nallapati, D. Zhang, Z. Wang, A. O. Arnold, and B. Xiang, "Improving factual consistency of abstractive summarization via question answering," in *Proc. 59th Annu. Meeting Assoc. Comput. Linguistics 11th Int. Joint Conf. Natural Lang. Process.*, 2021, pp. 6881–6894.
[13] S. Gehrmann, Y. Deng, and A. Rush, "Bottom-up abstractive summarization," in *Proc. Conf. Empirical Methods Natural Lang. Process.* Brussels, Belgium: Association for Computational Linguistics, Oct. 2018, pp. 4098–4109.
[14] J. Cho, M. Seo, and H. Hajishirzi, "Mixture content selection for diverse sequence generation," in *Proc. 2019 Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process. (EMNLP-IJCNLP)*. Hong Kong: Association for Computational Linguistics, Nov. 2019, pp. 3121–3131.
[15] X.-W. Han, H.-T. Zheng, J.-Y. Chen, and C.-Z. Zhao, "Diverse decoding for abstractive document summarization," *Appl. Sci.*, vol. 9, no. 3, p. 386, Jan. 2019. [Online]. Available: https://www.mdpi.com/2076-3417/9/3/386
[16] Z. Lin, Y. Gong, Y. Shen, T. Wu, Z. Fan, C. Lin, N. Duan, and W. Chen, "Text generation with diffusion language models: A pre-training approach with continuous paragraph denoise," in *Proc. 40th Int. Conf. Mach. Learn. (ICML)*, 2023, pp. 21051–21064.
[17] T. Wu, Z. Fan, X. Liu, Y. Gong, Y. Shen, J. Jiao, H.-T. Zheng, J. Li, Z. Wei, J. Guo, N. Duan, and W. Chen, "AR-Diffusion: Auto-regressive diffusion model for text generation," 2023, *arXiv:2305.09515*.
[18] S. Dathathri, A. Madotto, J. Lan, J. Hung, E. Frank, P. Molino, J. Yosinski, and R. Liu, "Plug and play language models: A simple approach to controlled text generation," in *Proc. Int. Conf. Learn. Represent.*, 2020.
[19] X. Li, J. Thickstun, I. Gulrajani, P. Liang, and T. B. Hashimoto, "Diffusion-LM improves controllable text generation," in *Proc. Adv. Neural Inf. Process. Syst.*, New Orleans, LA, USA, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds., Dec. 2022, pp. 4328–4343.
[20] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "BERTScore: Evaluating text generation with BERT," in *Proc. 8th Int. Conf. Learn. Represent.*, 2020.
[21] H. Touvron et al., "Llama 2: Open foundation and fine-tuned chat models," 2023, *arXiv:2307.09288*.
[22] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de L. Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M.-A. Lachaux, P. Stock, T. Le Scao, T. Lavril, T. Wang, T. Lacroix, and W. E. Sayed, "Mistral 7B," 2023, *arXiv:2310.06825*.
[23] J. Li, J. Liu, J. Ma, W. Yang, and D. Huang, "Boundary-aware abstractive summarization with entity-augmented attention for enhancing faithfulness," *ACM Trans. Asian Low-Resource Lang. Inf. Process.*, vol. 23, no. 4, pp. 1–18, Apr. 2024.
[24] M. Cao, Y. Dong, J. Wu, and J. C. K. Cheung, "Factual error correction for abstractive summarization models," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2020, pp. 6251–6258.
[25] Y. Dong, S. Wang, Z. Gan, Y. Cheng, J. C. K. Cheung, and J. Liu, "Multi-fact correction in abstractive text summarization," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2020, pp. 9320–9331.
[26] J. Shin, S.-B. Park, and H.-J. Song, "Token-level fact correction in abstractive summarization," *IEEE Access*, vol. 11, pp. 1934–1943, 2023.
[27] T. Dixit, F. Wang, and M. Chen, "Improving factuality of abstractive summarization without sacrificing summary quality," in *Proc. 61st Annu. Meeting Assoc. Comput. Linguistics*, A. Rogers, J. Boyd-Graber, and N. Okazaki, Eds. Toronto, ON, Canada: Association for Computational Linguistics, Jul. 2023, pp. 902–913.
[28] L. Qian, L. Qiu, W. Zhang, X. Jiang, and Y. Yu, "Exploring diverse expressions for paraphrase generation," in *Proc. Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process.* Hong Kong: Association for Computational Linguistics, Nov. 2019, pp. 3171–3180.
[29] Z. Lin, G. Indra Winata, P. Xu, Z. Liu, and P. Fung, "Variational transformers for diverse response generation," 2020, *arXiv:2003.12738*.
[30] Z. Shi, X. Chen, X. Qiu, and X. Huang, "Toward diverse text generation with inverse reinforcement learning," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, Jul. 2018, pp. 4361–4367.
[31] A. K. Vijayakumar, M. Cogswell, R. R. Selvaraju, Q. Sun, S. Lee, D. Crandall, and D. Batra, "Diverse beam search: Decoding diverse solutions from neural sequence models," 2016, *arXiv:1610.02424*.
[32] A. Fan, M. Lewis, and Y. Dauphin, "Hierarchical neural story generation," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*. Melbourne, VIC, Australia: Association for Computational Linguistics, 2018, pp. 889–898.
[33] A. Holtzman, J. Buys, L. Du, M. Forbes, and Y. Choi, "The curious case of neural text degeneration," in *Proc. Int. Conf. Learn. Represent.*, 2020.
[34] H. Cao, C. Tan, Z. Gao, Y. Xu, G. Chen, P.-A. Heng, and S. Z. Li, "A survey on generative diffusion models," *IEEE Trans. Knowl. Data Eng.*, vol. 36, no. 7, pp. 2814–2830, Jul. 2024.
[35] Q. Yi, X. Chen, C. Zhang, Z. Zhou, L. Zhu, and X. Kong, "Diffusion models in text generation: A survey," *PeerJ Comput. Sci.*, vol. 10, p. e1905, Feb. 2024.
[36] S. Gong, M. Li, J. Feng, Z. Wu, and L. Kong, "DiffuSeq: Sequence to sequence text generation with diffusion models," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2023, pp. 1–13.
[37] H. Yuan, Z. Yuan, C. Tan, F. Huang, and S. Huang, "SeqDiffuSeq: Text diffusion with encoder–decoder transformers," 2022, *arXiv:2212.10325*.
[38] J. Ye, Z. Zheng, Y. Bao, L. Qian, and M. Wang, "DINOISER: Diffused conditional sequence learning by manipulating noises," 2023, *arXiv:2302.10025*.
[39] J. Lovelace, V. Kishore, C. Wan, E. S. Shekhtman, and K. Q. Weinberger, "Latent diffusion for language generation," in *Proc. 37th Conf. Neural Inf. Process. Syst.*, 2023, pp. 1–28. [Online]. Available: https://openreview.net/forum?id=NKdtztladR
[40] Y. Zhang, J. Gu, Z. Wu, S. Zhai, J. M. Susskind, and N. Jaitly, "PLANNER: Generating diversified paragraph via latent language diffusion model," in *Proc. 37th Conf. Neural Inf. Process. Syst.*, 2023, pp. 80178–80190.
[41] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Proc. 34th Int. Conf. Neural Inf. Process. Syst.* Red Hook, NY, USA: Curran Associates Inc., 2020, pp. 6840–6851.
[42] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. U. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inform. Process. Syst. (NIPS)*, 2017, pp. 5998–6008.
[43] D. Wan and M. Bansal, "Evaluating and improving factuality in multimodal abstractive summarization," in *Proc. Conf. Empirical Methods Natural Lang. Process.* Abu Dhabi, UAE: Association for Computational Linguistics, Dec. 2022, pp. 9632–9648.

[44] S. Gabriel, A. Bosselut, J. Da, A. Holtzman, J. Buys, K. Lo, A. Celikyilmaz, and Y. Choi, "Discourse understanding and factual consistency in abstractive summarization," in *Proc. 16th Conf. Eur. Chapter Assoc. Comput. Linguistics, Main Volume*, Apr. 2021, pp. 435–447.
[45] S. Banerjee and A. Lavie, "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments," in *Proc. ACL Workshop Intrinsic Extrinsic Eval. Measures Mach. Transl. Summarization*. Ann Arbor, MI, Michigan: Association for Computational Linguistics, Jun. 2005, pp. 65–72.
[46] R. Vedantam, C. L. Zitnick, and D. Parikh, "CIDEr: Consensus-based image description evaluation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 4566–4575.
[47] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," in *Proc. Int. Conf. Learn. Represent.*, 2021.
[48] J. Yu, Y. Wang, C. Zhao, B. Ghanem, and J. Zhang, "FreeDoM: Training-free energy-guided conditional diffusion model," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 23174–23184.
[49] K. M. Hermann, T. Kočiský, E. Grefenstette, L. Espeholt, W. Kay, M. Suleyman, and P. Blunsom, "Teaching machines to read and comprehend," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 1–9.
[50] S. Narayan, S. B. Cohen, and M. Lapata, "Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2018, pp. 1797–1807.
[51] A. M. Rush, S. Chopra, and J. Weston, "A neural attention model for abstractive sentence summarization," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2015, pp. 379–389.
[52] Y. Liu, S. Feng, X. Han, V. Balachandran, C. Y. Park, S. Kumar, and Y. Tsvetkov, "P^3Sum: Preserving author's perspective in news summarization with diffusion language models," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol.*, K. Duh, H. Gomez, and S. Bethard, Eds. Mexico City, Mexico: Association for Computational Linguistics, 2024, pp. 2154–2173.
[53] S. Kumar and W. Byrne, "Minimum Bayes-risk decoding for statistical machine translation," in *Proc. Hum. Lang. Technol. Conf. North Amer. Chapter Assoc. Comput. Linguistics (HLT-NAACL)*. USA: Association for Computational Linguistics, 2004, pp. 169–176.
[54] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Proc. Text Summarization Branches Out*. Barcelona, Spain: Association for Computational Linguistics, Jul. 2004, pp. 74–81.
[55] A. Fabbri, C.-S. Wu, W. Liu, and C. Xiong, "QAFactEval: Improved QA-based factual consistency evaluation for summarization," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, M. Carpuat, M.-C. de Marneffe, and I. V. M. Ruiz, Eds. Seattle, WA, United States: Association for Computational Linguistics, Jul. 2022, pp. 2587–2601.
[56] Y. Zhu, S. Lu, L. Zheng, J. Guo, W. Zhang, J. Wang, and Y. Yu, "Texygen: A benchmarking platform for text generation models," in *Proc. 41st Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.* New York, NY, USA: Association for Computing Machinery, Jun. 2018, pp. 1097–1100.
[57] K. Owczarzak, J. M. Conroy, H. T. Dang, and A. Nenkova, "An assessment of the accuracy of automatic evaluation in summarization," in *Proc. Workshop Eval. Metrics Syst. Comparison Autom. Summarization*. Montral, QC, Canada: Association for Computational Linguistics, Jun. 2012, pp. 1–9.
[58] M. Desmond, Z. Ashktorab, Q. Pan, C. Dugan, and J. M. Johnson, "EvaluLLM: LLM assisted evaluation of generative outputs," in *Proc. Companion 29th Int. Conf. Intell. User Interfaces*. New York, NY, USA: Association for Computing Machinery, Mar. 2024, pp. 30–32, doi: 10.1145/3640544.3645216.
[59] Y. Liu, D. Iter, Y. Xu, S. Wang, R. Xu, and C. Zhu, "G-Eval: NLG evaluation using GPT-4 with better human alignment," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, H. Bouamor, J. Pino, and K. Bali, Eds. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 2511–2522.
[60] C.-H. Chiang and H.-Y. Lee, "Can large language models be an alternative to human evaluations?" in *Proc. 61st Annu. Meeting Assoc. Comput. Linguistics*. Toronto, ON, Canada: Association for Computational Linguistics, 2023, pp. 15607–15631.
[61] A. Sottana, B. Liang, K. Zou, and Z. Yuan, "Evaluation metrics in the era of GPT-4: Reliably evaluating large language models on sequence to sequence tasks," in *Proc. Conf. Empirical Methods Natural Lang. Process.* Singapore: Association for Computational Linguistics, 2023, pp. 8776–8788.
[62] Y. Song, P. Dhariwal, M. Chen, and I. Sutskever, "Consistency models," in *Proc. 40th Int. Conf. Mach. Learn.*, 2023, pp. 1–42.

JEONGWAN SHIN received the B.S. degree in information and communication engineering from Yeungnam University, in 2015, and the M.S. degree in computer science and engineering from Kyungpook National University, in 2018, where he is currently pursuing the Ph.D. degree in computer science and engineering. His research interests include machine learning, natural language processing, abstractive summarization, and fact correction.

HYEYOUNG PARK (Member, IEEE) received the B.S. (summa cum laude), M.S., and Ph.D. degrees in computer science from Yonsei University, Seoul, South Korea, in 1994, 1996, and 2000, respectively. She was a member of Research Staff with the Brain Science Institute, RIKEN, Japan, from 2000 to 2004. She is currently a Professor with the School of Computer Science and Engineering, Kyungpook National University, Daegu, South Korea. Her current research interests include computational learning theory and machine learning theory and their application to various fields, such as pattern recognition, image processing, and data mining.

HYUN-JE SONG (Member, IEEE) received the B.S. degree in computer engineering and the M.S. and Ph.D. degrees in computer science and engineering from Kyungpook National University, in 2008, 2010, and 2015, respectively. From 2016 to 2019, he was a Software Developer with Search and Clova Team, Naver Corporation. In 2019, he joined Jeonbuk National University, where he is currently an Associate Professor of computer science and artificial intelligence. His research interests include machine learning, natural language processing, information retrieval, and dialogue management.